ISCAM 2: CHAPTER 3 EXERCISES

1. Feeling Motivated?

A psychology study investigated whether people display more creativity when they are thinking about intrinsic or extrinsic motivations. The subjects were 47 people with extensive experience with creative writing. They were randomly assigned to one of two groups: one group answered a survey about intrinsic motivations for writing (such as the pleasure of self-expression) and the other group answered a survey about extrinsic motivations (such as public recognition). Then all subjects were instructed to write a Haiku poem, and these poems were evaluated for creativity by a panel of judges. The researchers conjectured that subjects who were thinking about intrinsic motivations would display more creativity than subjects who were thinking about extrinsic motivations. The creativity scores from this study are below and also in the file `creativity.txt`.

(a) Identify the explanatory and response variables. Also classify each as categorical or quantitative.
(b) Is this an observational study or a randomized experiment? Explain how you know.
(c) Examine the dotplots of the sample data produced by the Comparing Groups applet. Submit a screen capture of these graphs, and comment on what they reveal about the researchers' conjecture.
(d) Report the mean of the creativity scores for each group. Do these summary values indicate that the intrinsically motivated group did indeed display more creativity than the intrinsically motivated group?
(e) Carry out a randomization test using technology to the data provide statistically significant evidence that the type of motivation causes affects creativity score in the conjectured direction. Submit a screen capture of the resulting dotplot, and answer four questions:
    i) Describe the null model that underlies this simulation analysis.
    ii) Explain what variable is displayed in the dotplot.
    iii) Describe what the dotplot reveals.
    iv) Report the approximate p-value.
(f) Summarize your conclusion in the context of this study. Include an explanation of the reasoning process behind your conclusion. Be sure to address the issues of causation (i.e., is a cause-and-effect conclusion warranted?) and generalizability (i.e., how broadly can you legitimately generalize your conclusion?), as well as the issue of statistical significance.

2. Feeling Motivated? (cont.)

Reconsider the previous study.

(a) Suppose you thought the intrinsic motivation would, on average, add 10 points to the creativity scores. Specify the corresponding null and (two-sided) alternative hypotheses.
(b) Open the `creativity.txt` file. Are the data in stacked or unstacked format?
(c) Copy and paste the data into the flash-based Randomization Test applet. This applet lets you specify a hypothesized group 1 effect. Specify 10 as the hypothesized group 1 effect and generate 1000 repetitions. Explain why this distribution is centered where it is.
(d) Count the samples beyond the observed difference in sample means. Does 10 appear to be a plausible value for the difference in the underlying treatment means? Explain your reasoning.
(e) Use R or Minitab to compute a 95% confidence interval comparing the two groups. Include your output and interpret the interval.
(f) Using the confidence interval, does 10 appear to be a plausible value for the difference in the underlying treatment means? Explain your reasoning.

Extra Credit: Use R or Minitab to carry out the two-sample *t*-test to obtain a p-value.

3. Guess the Instructor's Age

The file AgeGuesses.txt contains guesses of an instructor's age by her current students.

Let μ represent the average guess of her age by all current at the university and suppose the sample constitutes a representative sample of all students at this school on this issue. Because there is just one variable and we are not comparing groups, a "one-sample *t*-interval" could be used. This procedure is valid as long as the population distribution is normal or the sample size is large (30 is often used as a cut-off for "large").

(a) Use technology to determine a 90% one-sample *t*-interval for these data.

Minitab	**R**
• Select **Stat >Basic Statistics > One-sample t** • Specify the column containing the data or determine and enter the relevant summary statistics. • Under **Options**, specify the confidence level to be 90%.	• For example: t.test(guesses, alt="two.sided", conf.level=.90)

Include your output.

(b) Count how many of the class guesses are inside the 90% confidence interval. Is this close to 90%? Should it be?

(c) Suppose the population mean guess of my age was μ =40 years with a population standard deviation of σ =5 years. Open the Simulating Confidence Intervals applet and use the pull-down menu to select Means. Specify these values for μ, σ, and the sample size from our study. Generate 1000 intervals (e.g., 200 at a time 5 times), what is the "running total" of intervals that capture the population mean?

(d) The default method used in (c) assumes the value of σ is known, but this is seldom the case. Use the second pull-down menu to specify "z with s." Generate 1000 intervals and report the running total. What is a key difference between these intervals and those generated with the "z with sigma" method?

(e) Now suppose the sample size had only been 5. Repeat (d) for this sample size and report the running total.

(f) Now use the second pull-down menu to select "t". This creates the one-sample *t*-confidence interval for each sample. Generate 1000 intervals and based on these results explain why this procedure (using the *t* critical instead of the *z* critical as in (e)) would be preferred for the small sample size.

(g) Instead of estimating the population mean, we often want to predict the next outcome. If we wanted to instead say something like "I think 90% of student guesses will be between these two numbers" we have to calculate a *prediction interval* instead of a *confidence interval*. The formula for a prediction interval is Investigation 3.3. Carry out the calculations (by hand) for a 90% prediction interval for a Cal Poly student's guess of her age.

(h) How does the prediction interval compare (e.g., midpoint, length) to the confidence interval?

4. Low Carb Diet

A study by Foster el al., reported in *The New England Journal of Medicine* (May, 2003), investigated the effectiveness of a popular "low-carb" diet. The researchers randomly assigned 63 obese men and women to either a low-carbohydrate, high-protein, high-fat (Atkins) diet or a low-calorie, high-carbohydrate, low-fat (conventional) diet. The mean amount of weight lost, as percent of body weight, after 3 months, 6 months and 12 months are shown in the table below.

(The baseline weight was carried forward in the case of missing values.)

Time	Diet	Sample size	Mean	SD
3 months	Low-carb	33	6.8	5.0
	Conventional	30	2.7	3.7

6 months	Low-carb	33	7.0	6.5
	Conventional	30	3.2	5.6
12 months	Low-carb	33	4.4	6.7
	Conventional	30	2.5	6.3

(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory and response variables.
(c) Report the relevant hypotheses (in symbols) for testing whether the mean weight losses differ significantly between the two diets.
(d) Calculate the *t*-test statistic for testing these hypotheses at the 3-month point. (You can use either a pooled or an unpooled test, but indicate which you use. Feel free to use R or Minitab or the Theory Based Inference applet or you may do this by hand.) Also report the p-value and your test decision at the .05 significance level.
(e) Repeat (d) for comparing the weight losses between the two diets at the 6-month point and again at the 12-month point.
(f) Summarize your conclusions from these three tests. In particular, what do you notice about the trend in the p-value as time passes, and what does that reveal?
(g) Report the 95% confidence intervals for the difference in mean weight loss between the two diets at each time point. (Again feel free to use software.) Comment on how these confidence intervals change across the three time points.

5. Marriage Ages

A student investigated whether husbands tend to be older than their wives. He gathered data on the ages of a sample of 24 couples, taken from marriage licenses filed in Cumberland County, Pennsylvania, in June and July of 1993. These data can be accessed in a file `MarriageAges.txt`.

(a) For each couple, calculate the difference in ages (taking the husband's age minus the wife's age). Produce and comment on a dotplot of these differences, keeping in mind the research question of whether husbands tend to be older than their wives.
(b) State the null and alternative hypotheses (in symbols) for testing whether the sample data support the research conjecture that husbands tend to be older than their wives.
(c) Copy/paste the data into the Matched Pairs Randomization applet, and perform 1000 repetitions of the randomization. Submit a copy of the resulting dotplot of sample mean differences. Also use the simulation results to determine an empirical p-value.
(d) Describe what the empirical p-value in (c) represents (it's the probability of what?), and summarize the conclusion that you draw from it.
(e) Investigate and comment on whether the technical conditions of a paired *t*-test appear to be satisfied here.
(f) Calculate the paired *t*-test statistic and p-value. Would you reject the null hypothesis at the .05 significance level?
(g) Produce and interpret a 90% confidence interval for the population mean difference in ages between a husband and wife.
(h) Produce and interpret a 90% prediction interval for the difference in age between a husband and wife.

6. Cool Mice

Medical examiners can use the temperature of a dead body at a murder scene to estimate the time of death. But can a clever murderer disguise the time of death by reheating the victim's body? A scientist actually investigated this issue on mice. Hart (1951) used 19 mice as the experimental units. He sacrificed each mouse and then measured the cooling constant of its body. Then he reheated the mouse's body and measured its cooling constant in that reheated state. The results are in `CoolMice.txt`.

(a) Explain why these data call for a matched pairs analysis.
(b) Produce and comment on relevant graphical displays and numerical summaries for investigating the question of whether cooling constants for reheated mice are similar to those of freshly killed mice.
(c) Conduct a paired *t*-test or use the Matched Pairs Randomization applet to determine whether the data suggest a significant difference in average cooling constants between freshly killed and reheated mice. If you use the *t*-test, make sure comment on whether you believe the test procedure is valid and how you are decided.
(d) Construct and interpret a 95% confidence interval for estimating the population mean difference in cooling constants.
(e) Summarize the conclusions you would draw from this study. Make sure you comment on significance, confidence, generalizability, and causation.

7. Bumpus Data

In a famous 1898 lecture described in *The Statistical Sleuth*, a biologist named Bumpus presented data that he analyzed to study the process of natural selection. The data were obtained from adult male house sparrows, some of which had survived a particularly severe winter storm, and others of which had perished. Bumpus investigated whether those that survived had physical characteristics that may have helped them to withstand the storm. Data on the humerus (arm bone) lengths (in thousandths of an inch) follow and appear in `Bumpus.txt`:

Survived:
687 703 709 715 728 721 729 723 728 723 726 728 736 733 730 733 730 739 735
741 741 749 741 743 741 752 752 751 756 755 766 767 769 770 780

Perished:
659 689 703 702 709 713 720 729 726 726 720 737 739 731 738 736 738 744 745
743 754 752 752 765

(a) Is this an observational study or an experiment? Explain.
(b) Identify and classify the two variables represented in these data.
(c) Produce graphical and numerical summaries for comparing the distributions of humerus lengths between the two groups of sparrows. Write a paragraph addressing Bumpus' question of whether sparrows who survived tended to be physically superior (as measured by humerus length) to those who perished.

8. Bumpus Data (cont.)

Reconsider the previous question. Bumpus also recorded the weights (in grams) of each sparrow. One hypothesis is that heavier birds are bigger and stronger, therefore more likely to survive the storm. Another hypothesis is that heavier birds are less agile and less mobile, therefore less likely to survive the storm. A third possibility is that there is no association between a bird's weight and its capacity to survive the storm.

(a) Before you analyze the data, identify which of these three hypotheses you consider the most reasonable (intuitively). Explain briefly.

The data follow and appear in `Bumpus.txt`:

Survived:
24.5 26.9 26.9 24.3 24.1 26.5 24.6 24.2 23.6 26.2 26.2 24.8 25.4 23.7 25.7 25.7 26.3
26.7 23.9 24.7 28.0 27.9 25.9 25.7 26.6 23.2 25.7 26.3 24.3 26.7 24.9 23.8 25.6 27.0
24.7

Perished:
26.5 26.1 25.6 25.9 25.5 27.6 25.8 24.9 26.0 26.5 26.0 27.1 25.1 26.0 25.6 25.0 24.6
25.0 26.0 28.3 24.6 27.5 31.1 28.3

(b) Analyze these data with graphical and numerical summaries. Write a paragraph summarizing what your analysis reveals relevant to the competing hypotheses described above.

8.5 July Temperatures

The July 8, 2012 edition of the *San Luis Obispo Tribune* listed predicted high temperatures (in degrees Fahrenheit) for that date. One section reported predictions for locations in San Luis Obispo county, another section for locations throughout the state of California, and another section for cities across the United States. The data can be found in the file JulyTemps.txt.

(a) Produce (and submit) dotplots of the predicted high temperatures for the three regions, using the same scale and on the same axis for each dotplot.
(b) Calculate (and report) the mean and median, SD, and IQR of the temperatures for each region.
(c) Based on the graphs and statistics, write a paragraph comparing and contrasting the distributions of predicted high temperatures in the three regions. [*Hint*: As always when describing distributions of quantitative data, be sure to comment on center, variability, shape, and outliers.]
(d) Produce (and submit) histograms of the predicted high temperatures for the three regions, using the same scale for each histogram.
(e) The San Luis Obispo county and California region display some *bi-modality* in their distributions. Describe what this means, and provide an explanation for why it makes sense that these distributions reveal some bi-modality.
(f) Calculate (and report) the five-number summary of the temperatures for each region.
(g) Produce (and submit) boxplots of the predicted high temperatures for the three regions, using the same scale and on the same axis for each boxplot.
(h) Identify the location/city for any outliers revealed in the boxplots. Also use the 1.5×IQR criterion to verify (by hand) that the location/city really is an outlier.
(i) Now change the measurement units to be degrees Celsius rather than degrees Fahrenheit. [*Hint*: Create a new variable by first subtracting 32 from the temperature and then multiplying by 5/9.] Produce (and submit) dotplots of the predicted high temperatures (in degrees Celsius) for the three regions, using the same scale and on the same axis. Comment on how the shapes in these dotplots compare to the original dotplots (when the measurement units were degrees Fahrenheit).
(j) Calculate (and report) the mean and median, SD, and IQR of the temperatures (in degrees Celsius) for each region.
(k) Determine (and describe) how the values of these statistics have changed based on the transformation from degrees Fahrenheit to degrees Celsius. [*Hint*: Be as specific as you can be. For example, do not just say that the SD got smaller.]

9. 2004 U.S. Open

A tennis fan recorded data on a random sample of 16 first-round men's singles matches from the 2004 U.S. Open and also on a random sample of 16 first-round women's matches. (The fan did not want to invest the time required to gather and record the data for all matches played in the tournament.) Variables recorded include gender, number of sets played, number of games played, number of points played, and length of match in minutes.

(a) Classify each of these variables as categorical or quantitative.

The sorted data for the number of points played in a match are given here:

Men:	55	173	184	206	208	211	223	225	230	234	234	260	261	276	278	296
Women:	88	89	95	96	98	107	118	132	140	157	159	171	179	179	183	228

(b) Determine (by hand) the five-number summary for each gender's distribution of the number of points played by each gender.
(c) For each gender, determine whether there are any outliers by the 1.5IQR criterion (Investigation 3.1).

(d) Construct a boxplot for each gender's distribution, placing them on the same scale. (Remember to label you axes and include scales.)
(e) Comment on what the numerical and graphical summaries reveal about the distributions of points between the two genders.
(f) Did all of these men's matches play more points than all of the women's matches? Do men tend to play more points in their matches than women? Explain the difference in these two questions as you justify your answers.

10. 2004 U.S. Open (cont.)
Reconsider the tennis data from the 2004 U.S. Open.
(a) Before turning to technology, make (educated) guesses for the values of the mean and standard deviation of the number of points played for each gender. Briefly explain your guesses.
(b) Use technology (`USOpen04.txt`) to calculate these means and standard deviations. How were your guesses?
(c) The outlier is a men's match in which one player suffered an injury and had to retire early.
(d) Make predictions for the effect that removing the outlier would have on the mean, median, standard deviation, and IQR of the points played by men.
(e) Remove the outlier and re-calculate these statistics. Which statistics were more affected by the removal of the outlier? Explain why this makes sense.

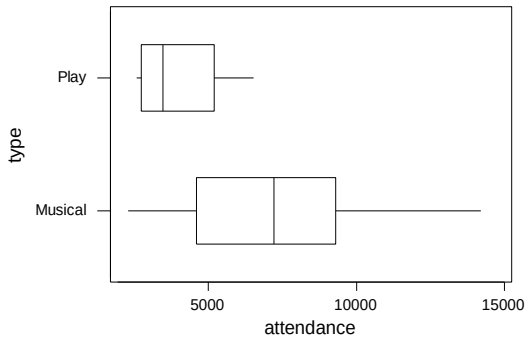
11. 2004 U.S. Open (cont.)
Reconsider the 2004 U.S. Open tennis data again (`USOpen04.txt`). Use technology to analyze the men's and women's distributions of the sets, games, and time variables. For each of these three variables, produce graphical and numerical summaries to compare the distributions between the two genders, and write a paragraph comparing and contrasting them.

12. 2004 U.S. Open (cont.)
Reconsider the 2004 U.S. Open tennis data yet again (`USOpen04.txt`). Use technology to create three new variables:

- Ratio of games to sets
- Ratio of points to games
- Ratio of time to points

Analyze these data to investigate whether men and women differ with regard to the distributions of these variables. For each of these three variables, produce graphical and numerical summaries to compare the distributions between the two genders, and write a paragraph comparing and contrasting them.

13. Broadway Attendance
The boxplots shown reveal the distributions of weekly attendance for Broadway shows in the first week of September in 1999, where the shows have been categorized as "play" or "musical."
(a) Did one type of show (play or musical) tend to have more attendees? Justify your conclusion.



(b) Did one type of show tend to have more variability in their attendance figures? Justify your conclusion.
(c) Which distribution appears to be more skewed? Explain how you are deciding.
(d) For the musicals, the mean was equal to 7121 and the standard deviation was equal to 3126. What are the "measurement units" of these numbers?
(e) For the musicals, between what two values do you expect to find the middle 68% of the attendance figures? Explain.

14. Memorizing Letters

Students in a statistics course at Cal Poly were given 20 seconds to memorize as many letters as possible in a sequence of 30 letters. The letters and the sequence were exactly the same for all students, but the presentation of the letters differed. Twenty-seven students were randomly assigned to see letters arranged in recognizable three-letter chunks such as JFK-CIA-FBI and so on. For the other 26 students, the letters were in less recognizable chunks such as JFKC-IAF and so on. Students' "scores" were determined as the number of letters they memorized correctly in the sequence before their first mistake.

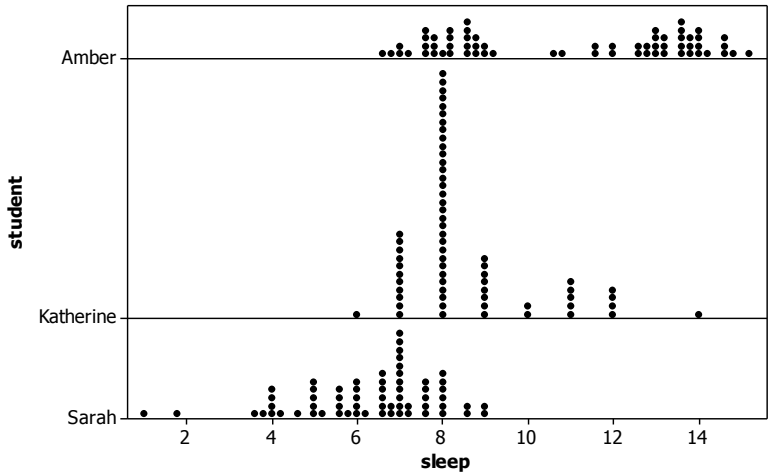
(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory and response variable. Identify each as categorical or quantitative.
(c) Which group would you expect to memorize more letters in general?

The resulting numbers of letters memorized successfully (`MemoryLetters.txt`) were:
JFK: 6, 6, 6, 8, 9, 9, 9, 9, 12, 15, 15, 15, 15, 18, 18, 18, 19, 21, 21, 21, 21, 21, 21, 21, 24, 27, 27
JFKC: 2, 3, 3, 3, 5, 6, 6, 6, 6, 8, 9, 9, 10, 13, 14, 14, 14, 14, 14, 15, 15, 15, 17, 18, 20, 24

(d) What proportion of the 27 scores in the JFK group are multiples of three? What about in the JFKC group of 26 scores? Explain why it makes sense that so many scores in the JFK group are multiples of three. (This aspect of a distribution, where the data are clustered at certain values, is called *granularity*.)
(e) Construct visual displays to compare the distributions of letters memorized correctly between the two groups. Report the five-number summary, as well as the mean and standard deviation, for each group. Write a paragraph comparing and contrasting the distributions. (Remember to comment on center, spread, shape, and outliers.)

15. Sleeping Students

The following dotplots display the distribution of sleeping times (per day, in hours) of three college students (Amber, Katherine, Sarah) for a nine-week period in the fall of 2004.



(a) One of these students developed mononucleosis during the term and so was told to get as much rest as possible for several weeks. Which student do you think this is? Explain your reasoning.
(b) One of these students is the mother of two small children. Which student do you think this is? Explain your reasoning.
(c) Which student recorded her sleeping times only to the nearest hour? Explain.
(d) Which student generally got the most sleep? Which generally got the least?
(e) For one of these students, her mean sleeping time exceeded her median sleeping time. Which student do you think this is? Explain your reasoning.

16. Sleeping Students (cont.)

Reconsider the students' sleeping times from the previous exercise. The data are in the worksheet `SleepStudents.txt`.

(a) Determine the five-number summary of sleeping times for each student.
(b) For each student, determine which (if any) of their sleeping times qualify as outliers by the 1.5IQR rule.
(c) Create boxplots of these students' sleeping times on the same scale. Comment on what these boxplots reveal.
(d) What does the dotplot reveal about Amber's sleeping times that the boxplot does not?

17. Sleeping Students (cont.)

Reconsider the students' sleeping times from the previous exercises (`SleepStudents.txt`).

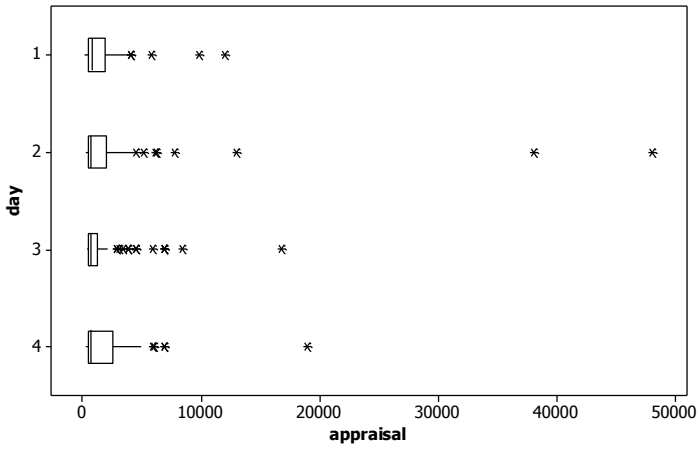
(a) Calculate the mean and standard deviation of sleeping times for each student.
(b) For each student, determine the proportion of the 63 sleeping times that fall within one standard deviation of the mean.
(c) For which student does the empirical rule appear to hold most closely? For that student, determine the proportion of sleeping times that fall within two standard deviations of the mean.
(d) Suppose that Katherine gets 10 hours of sleep in a particular night. How many hours more than her mean is this? Also calculate the *z*-score for this value.
(e) Suppose that Amber gets 13 hours of sleep in a particular night. How many hours more than her mean is this? Also calculate the *z*-score for this value.
(f) Which of these (10 hours for Katherine or 13 for Amber) is higher above that student's mean? Which has the higher z-score? Explain why your answers are not the same.

18. Sleeping Students (cont.)
Reconsider the students' sleeping times from the previous exercises (`SleepStudents.txt`). The worksheet also includes a day-of-the-week variable and a variable called *school night*? indicating whether school was in session the next day. For each student, analyze her sleeping times on school nights vs. non-school nights. Write a paragraph summarizing your findings. Also identify which student appears to have the biggest difference in sleeping times between these two kinds of days, and identify which has the least difference.

19. Surfboard Lengths
A student collected data on surfers over several weeks at a local beach (Wood, 2004). The data are in the file `surfer.txt`. Two of the questions of interest are how the age distributions of men and women surfers compare, and how the lengths of surfboards used by men and women compare.
(a) Identify the observational units in this study.
(b) Classify each of these variables (age, gender, surfboard length) as categorical or quantitative.
(c) Produce graphical displays and numerical summaries to address the question of how the age distributions of men and women surfers compare. Write a paragraph summarizing your findings. Include well-labeled output as appropriate.
(d) Produce graphical displays and numerical summaries to address the question of how the surfboard length distributions of men and women surfers compare. Write a paragraph summarizing your findings. Include well-labeled output as appropriate.

20. Health Club Ages
A student collected data on ages of people who joined a local health club in August and September of 2004, also recording the gender of each person (Schmitt, 2004). The student took a systematic sample of people who joined the club in August and an independent systematic sample of people who joined the club in September. The student wanted to compare the distributions of ages between males and females and also between new members who joined in August and September. The data are in the file `GymMembership.txt`. Analyze the data with appropriate graphical and numerical summaries, and write a 1-2-paragraph summary of your findings.

21. Appraisal Prices
The following boxplots are the appraisal prices of pieces of art auctioned off over a four-day period in December of 2004:



(a) Comment on what these four distributions have in common.
(b) Would you expect the mean appraisal price to be larger than, smaller than, or close to the median appraisal price on these days? Explain.
(c) Day 2 has the smallest median appraisal price among these four days, but it has the largest mean. Explain, based on the boxplots, why this makes sense.

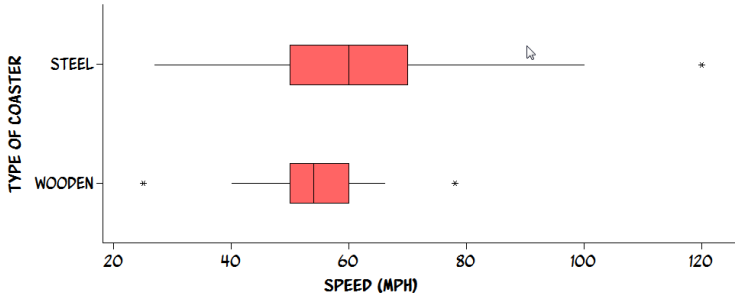
22. Appraisal Prices (cont.)

The auction data from the previous exercise appear in `auction.txt`, where the variables are day, appraisal price, starting price at the auction, and selling price at the auction.

(a) Create a new variable: ratio of starting price to appraisal price. How many and what proportion of the art pieces had a starting price of more than half their appraisal price? How many and what proportion of the art pieces had a starting price less than one-third their appraisal price?
(b) Produce graphical displays and numerical summaries to analyze the distribution of this "ratio" variable. Write a paragraph reporting your findings.
(c) Now compare the distribution of these ratios across the four days of the auction. Do the distributions appear to differ considerably across the days? Write a paragraph reporting your findings.

23. Roller Coaster Speeds

The Roller Coaster Database maintains a website (www.rcdb.com) with data on roller coasters around the world. Some of the data recorded include whether the coaster is made of wood or steel and the maximum speed achieved by the coaster, in mile per hour. The boxplots shown display the distributions of speed by type of coaster for 145 coasters in the United States as downloaded from the site in November of 2003.



(a) Do these boxplots allow you to determine whether there are more wooden or steel roller coasters?
(b) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 60 mph? Explain, and if so, answer the question.
(c) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 50 mph? Explain, and if so, answer the question.
(d) Do these boxplots allow you to say which type has a higher percentage of coasters that go faster than 48 mph? [*Hint*: Think twice on this one.]
(e) The steel coasters have a "high outlier." Explain how I know this from the above display.
(f) Conjecture as to how the mean, median, interquartile range, and standard deviation will change (if at all) if that coaster identified in part (e) (Top Thrill Dragster in Cedar Point Amusement Park, Sandusky, Ohio) is removed from the data set. Explain your reasoning.

24. Roller Coaster Speeds (cont.)

Reconsider the data in the previous exercise on 139 coasters in the United States, as downloaded from the www.rcdb.com site in November of 2003 (`coasters.txt`).

(a) Identify the observational units in this study. Then identify the explanatory and the response variable

here. Also indicate for each whether it is quantitative or categorical.

(b) Write a paragraph comparing and contrasting these distributions. Describe the shaper, center, and spread (as best you can) for each distribution, and then also comment on the issue of whether one type of coaster tends to have higher speeds than the other. Remember to state your description in the context of the study.

25. Roller Coaster Speeds (cont.)

(a) Open the data file `coasters.txt`, which contains data on 145 roller coasters in the United States, as downloaded from the www.rcdb.com site in November of 2003. Use technology to produce boxplots of height (in feet) by type, length (in feet) by type, and drop (in feet) by type. Write a paragraph summarizing differences between wooden and steel coasters with regard to these variables.

(b) Another variable in the file is *age group* (column 13) which is coded as "1:older" for coasters opened in 1990 or earlier, coded as "2:middle" for coasters opened between 1991 and 1998 inclusive, and coded as "3:newer" for coasters opened in 1999 or later. Produce boxplots of height, length, drop and speed by this *age group* variable. Write a paragraph summarizing how roller coasters appear to have changed over time with respect to these variables.

26. Hypothetical Quiz Scores

Reconsider the hypothetical quiz scores for classes A–D in Practice Problem 3.1B.

(a) For each class (A–D), calculate the range of the quiz scores.

(b) Is the range a helpful measure here is comparing the variability of these distributions? Explain.

27. Create an Example

(a) Create a hypothetical example of 10 exam scores (say, between 0 and 100 with repeats allowed) such that 90% of the scores are above the mean.

(b) Repeat (a) for the condition that the mean is roughly 40 points less than the median.

(c) Repeat (a) for the condition that the IQR equals 0 and the mean is more than twice the median.

28. Measures of Center and Spread

The *mid-range* of a dataset is defined to be the sum of the minimum and maximum values divided by 2.
The *mid-hinge* of a dataset is defined to be the sum of the first and third quartiles divided by 2.

(a) Is mid-range a measure of center or a measure of spread? Explain.

(b) Is mid-hinge a measure of center or a measure of spread? Explain.

(c) Is the mid-range resistant to outliers? Explain.

(d) Is the mid-hinge resistant to outliers? Explain.

29. Identifying Outliers

Perhaps you are wondering about the motivation behind the "1.5IQR criterion" for identifying outliers.

(a) Determine the 25^{th} and 75^{th} percentiles of the standard normal model. Then calculate the inter-quartile range. Also draw a well-labeled sketch of the standard normal curve and indicate how to find the value of the IQR on the graph.

(b) Using the "1.5IQR" rule for identifying outliers, determine what proportion of the values from a standard normal distribution would be classified as outliers. [*Hint*: Again draw a sketch first, and then identify the "cut-off" points for identifying outliers using your answers from (a).]

(c) Use a simulation as a check on your calculations: First simulate 1000 random values from a standard

normal distribution. Then determine the IQR for your 1000 simulated values. Finally, set up an indicator variable to count how many of the values are not outliers. Also draw a boxplot to reveal the outliers. What proportion of the 1000 random values are identified as outliers? Is this close to your answer to (b)?

(d) Now consider a more general normal model with mean μ and standard deviation σ. Determine how your answers to (a) and (b) will change, if it all. Follow up with a technology simulation using a few different values of (μ, σ) as a check on your work. Summarize your results.

(e) Based on your simulation in (c), what proportion of the 1000 random values are more than 1IQR from the respective quartiles? What proportion of the 1000 random values are more than 2IQR from the respective quartiles? Explain why someone might consider 1.5IQR a more reasonable way to identify outliers than 1IQR or 2IQR.

(f) The rule of "3IQR" has also been recommended as a way to identify "extreme" outliers. What proportion of your simulated values are more than 3IQR are from the quartiles?

30. Identifying Outliers (cont.)

Reconsider the previous question. An alternative procedure for identifying outliers is to classify any value more than three standard deviations away from the mean as an outlier.

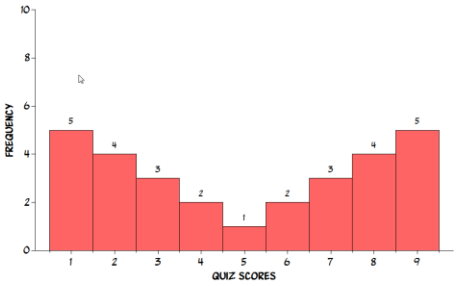
(a) By this criterion, what proportion of values from a normal distribution will be identified as outliers? Is this more or less than with the 1.5IQR criterion? Much more so?

(b) Repeat (a) if the criterion is to classify any observation more than *two* standard deviations away from the mean as an outlier.

(c) Explain how the 1.5IQR rule is a more "general" criterion than using 2 or 3 standard deviations? [*Hint*: When would the latter condition not be reasonable to apply?]

31. Properties of Center and Spread

The following histogram displays the (hypothetical) quiz scores for a class of $n = 29$ students.



Suppose we were to give every student 5 bonus points.

(a) How would the mean change? The median?

(b) How would the standard deviation change? The inter-quartile range?

Note: You should explain your answers to (a) and (b) without carrying out the calculations to find these new values.

32. Linear Transformations

Suppose that a *linear* transformation is applied to a set of data, so all of the x_i's are converted into y_i's by the expression $y_i = a + b\,x_i$ for some constants a and b. It can be shown that the mean of the transformed data is $\bar{y} = a + b\bar{x}$ and the standard deviation is SD(y) = bSD(x).

(a) Prove these results (using summation notation).

(b) Determine the effect of this linear transformation on the *median* of the data? Justify your answer. Prove that your answer is correct, making sure you thoroughly explain your proof.

(c) Determine the effect of this linear transformation on the *IQR* of the data? Justify your answer. Prove that your answer is correct, making sure you thoroughly explain your proof.

33. Seeding Clouds

Reconsider the cloud seeding data (CloudSeeding.txt) from Investigation 3.9 where you found the mean rainfall amount was 164.6 acre-feet for the unseeded clouds and 442.0 acre-feet for the seeded clouds.

(a) Use technology to take the (natural) log transformation of the rainfall amounts. Calculate and report the mean and median of these transformed values.
(b) Does the mean of the ln(rainfall) amounts equal the ln of the mean of the rainfall amounts? Report calculations to support your answer.
(c) Does the median of the ln(rainfall) amounts equal the ln of the median of the rainfall amounts? Report calculations to support your answer.
(d) Will the relationship that you found in (c) always hold? If so, explain. If not, provide a counterexample.

34. Log Transformations

Suppose that a *logarithmic* transformation is applied to a set of data, so all of the x_i's are converted into y_i's by the expression $y_i = \log(x_i)$.

(a) Explain why you cannot say what effect this would have on the mean of the data.
(b) Describe what effect this would have on the median of the data, and justify your answer.
(c) Between the IQR and standard deviation, for which measure can you say what the effect would be? Describe that effect, and justify your answer.

35. Seeding Clouds (cont.)

Reconsider the cloud seeding data (CloudSeeding.txt). At the end of Investigation 3.9, you applied the log transformation to the rainfall amounts.

(a) Use technology to take the square root of the rainfall amounts. Produce graphical and numerical summaries for comparing the two groups on this transformed variable. Comment on what your analysis reveals.
(b) Repeat (a) for the reciprocal transformation.
(c) Which of the three transformations that you have tried thus far (log, square root, reciprocal) does the best job of making the distributions more symmetric? Justify your choice.

36. Transformations

Consider a general power transformation, represented by the function $f(x) = x^p$, for some power p.

(a) Explain why using the power $p = 0$ does not make sense.
(a) The log transformation actually "takes the place" of zero on the power transformation scale. You can see this by examining derivatives.
(b) Take the derivative (with respect to x, for a fixed value of p) of $f_p(x) = x^p$.
(c) Take the derivative of $f(x) = \log(x)$.
(d) Explain how these derivatives reveal that $\log(x)$ is comparable to a power of zero on the power transformation scale. [*Hint*: $f'(x)$ has the same exponent on x as $f_p'(x)$ for what value of p?]

37. Body Mass Index

The data in BodyMassIndex.txt are ages (in years), weights (in kg), and heights (in cm) for a sample of adults (Heinz et al., 2003). Body mass index (BMI) is defined to be a person's weight (in kg) divided by the square of their height (in meters).

(a) Use technology to calculate the BMI values for this sample of adults by computing

$$BMI = (weight)/(height)^2 \times 1000.$$

(a) Produce boxplots and descriptive statistics comparing BMI values between men and women. Write a paragraph summarizing your findings. [Remember to comment on center, spread, and shape.]

(b) Try several transformations (log, square root, reciprocal) of the BMI values for the two genders combined. Identify which transformation produces an approximately symmetric distribution for the BMI values. Provide graphical displays to support your answer.

(c) Examine histograms of the BMI values for men and women separately. Then repeat this transformation analysis for men and for women separately. For each gender, identify which transformation produces an approximately symmetric distribution for the BMI values. Provide graphical displays to support your answer.

38. Mean IQs

Is it possible for an individual to move from one city to another and have the mean IQ decrease in both cities? If not, explain why not. If so, explain what conditions would be needed to make this happen.

39. Average Children

Suppose that you record the number of children in each of ten families (labeled as A–J) to be:

Family	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Number of children	1	2	1	0	2	2	3	7	4	2

(a) Determine the average (mean) number of children per family.

Now consider the 24 children in these families as the observational units, and consider the variable "number of siblings." Thus, the one child in family A has 0 siblings, each of the two children in family B has 1 sibling, and so on.

(b) Determine the average number of siblings per child.

(c) Some might expect that there would be a clear relationship between these two averages. For example, some might suspect that the average number of siblings would be one less than the average number of children. Give a mathematical explanation for why this is not the case.

40. Average Children (cont.)

Reconsider the previous question. A similar phenomenon can reveal itself with class sizes. The average number of students per class can be very different from the average class size per student. Demonstrate this with a hypothetical example of five classes. Specify the number of students in each class, and then calculate the average number of students per class. Then consider the students as the observational units, with "number of students in that student's class" as the variable, and calculate the average class size per student. Construct your example so that these two averages are quite different, and explain why that happens.

41. Body Mass Index (cont.)

Suppose that the body mass index (BMI) of healthy American males follows a symmetric, mound-shaped distribution with mean 24.5 and standard deviation 3.0 and that the BMI of healthy American females follows a symmetric, mound-shaped distribution with mean 22.5 and standard deviation 3.0.

(a) Between what two values would approximately 95% of males' BMI values fall?
(b) About what percentage of male BMI values fall below 21.5?
(c) About what percentage of male BMI values fall above 30.5?
(d) About what percentage of female BMI values fall between 19.5 and 25.5?
(e) About what percentage of female BMI values fall between 16.5 and 25.5?
(f) Below what value do about 2.5% of female BMI values fall?

42. SATs

Suppose the distribution of SAT scores is mound-shaped and symmetric with a mean of 1500 and a standard deviation of 240, and that the distribution of ACT scores is mound-shaped and symmetric with a mean of 21 and a standard deviation of 5. Suppose Tory scores a 1800 on the SATs and Jeff scores a 28 on the ACT.

(a) Provide a rough sketch, labeling the horizontal axis, of each distribution and indicate where the observed test score falls on the distribution.
(b) Which test taker had a higher score relative to the distribution of scores on that test? Explain. [*Hint*: Compare their *z*-scores.]

43. SATs (cont.)

Recall the previous Exercise, in which you considered SAT scores and ACT scores to have symmetric, mound-shaped distributions. Continue to assume that SAT scores have mean 1500 and standard deviation 240, while ACT scores have mean 21 and standard deviation 5.

(a) An ACT score of 21 is equivalent to what SAT score, in terms of z-scores?
(b) An ACT score of 26 is equivalent to what SAT score, in terms of z-scores?
(c) An ACT score of 28 is equivalent to what SAT score, in terms of z-scores?
(d) Let x represent a generic ACT score, and let y represent the SAT score to which x is equivalent, in terms of z-scores. Determine y as a function of x.
(e) Graph the function in (d), and confirm that it satisfies your answers to (a), (b), and (c).

44. Equating *z*-scores

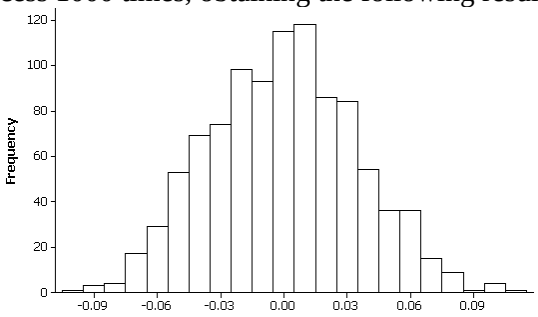
Reconsider the previous exercise. Suppose that two variables both have symmetric, mound-shaped distributions, and you want to find the value of one variable (call it y) that has the same z-score as a given value of the other variable (call it x). Denote the means of the variables by μ_x and μ_y, and denote their standard deviations by σ_x and σ_y.

(a) Derive a function that expresses y as a function of x, μ_x, μ_y, σ_x, and σ_y.
(b) If all else remains unchanged, is y an increasing or a decreasing function of x? Explain both algebraically and intuitively.
(c) Repeat (b), answering whether y is an increasing or a decreasing function of μ_x.
(d) Repeat (b), answering whether y is an increasing or a decreasing function of μ_y.
(e) Repeat (b), answering whether y is an increasing or a decreasing function of σ_x.
(f) Repeat (b), answering whether y is an increasing or a decreasing function of σ_y.

45. Normal Groceries

Suppose you take a random sample of 30 grocery products from two local stores and find that average price difference in these products is \$0.10, with standard deviation \$0.20. To decide if this is a statistically significant average price difference, suppose you simulate selecting random samples of 30 products from a normal distribution with mean 0 and standard deviation of 0.20, compute the sample

mean, and then repeat this process 1000 times, obtaining the following results.



(a) Specify the observational units in this graph and provide an appropriate label for the horizontal axis.
(b) Use the Central Limit Theorem to determine the theoretical standard deviation of this distribution. Does your result seem consistent with the above graph? Explain.
(c) Using these simulation results, would you consider \$0.10 a surprising average price difference if the population mean price difference was zero? Explain.
(d) What conclusion would you come to about the average price difference of all the products in these two stores? Explain.
(e) What part, if any, of the above analysis depends on the population following a normal distribution? Explain.

46. Exponential Models

Consider the exponential model, with probability density function $f(x) = (1/\beta)e^{-(x/\beta)}$ for $x > 0$. First consider the model with $\beta = 1$.

(a) Write out and sketch the pdf for this model with $\beta = 1$.

Another function that can be used to describe a probability model is a cumulative distribution function (cdf). The cdf is denoted by $F(x)$ and is defined to be the function that reports the probability that the random variable is less than or equal to the input of the function: $F(x) = P(X \leq x)$.

(b) Determine and sketch a well-labeled graph of the cdf of the exponential model with $\beta = 1$. [*Hint*: What is the functional form of $P(X \leq x)$ for all values of x?]

The median of a continuous probability model is defined to be a value m such that $P(X \leq m) = 0.5$ and $P(X \geq m) = 0.5$.

(c) Use the cdf to determine the median of the exponential model with $\lambda = 1$. [*Hint*: Set $F(m) = 0.5$ and solve for m.]

The mean, or **expected value**, of a continuous probability model, denoted as either E(X) or μ, is defined by $\mu = -\int_{-\infty}^{\infty} xf(x)dx$, where $f(x)$ is the probability density function.

(d) Verify that the mean of the exponential model with $\beta = 1$ is 1. [*Hint*: Use integration by parts.]
(e) How does the median compare to the mean for this exponential model? Explain why this makes sense, based on the shape of the density function.
(f) Use a Minitab or R simulation to verify your results. First simulate 1000 values from this exponential model.

Minitab	**R**

`MTB> rand 1000 c1;`
`SUBC> expo 1.`
or select **Calc > Random Data > Exponential** and ask for 1000 rows in C1 with a scale parameter of 1 and a threshold parameter of 0.
Note: scale = mean
	`> mydata = rexp(n=1000, rate = 1)`
	Note: rate = 1/mean

Then examine a histogram of the generated values and calculate descriptive statistics. Does the histogram follow the same shape as the density function? Do the median and mean values come close to your theoretical analysis?

47. Exponential Models (cont.)

Reconsider the previous question about the exponential probability model with parameter $\beta = 1$. Now consider the general exponential model with parameter β.

(a) Determine and sketch a well-labeled graph of the cumulative distribution function.
(b) Determine the median.
(c) Verify that the mean equals the parameter β.
(d) How do the mean and median compare?
(e) Show that the ratio of mean to median is constant regardless of β.
(f) Choose two different values of β (other than 1), and use a simulation to verify your findings. (Include a histogram and descriptive statistics of your generated distributions.)

48. Probability Density Functions

Consider the probability density function (model) for a random variable X given by

$$f(x) = (1+\theta x)/2 \text{ for } -1 < x < 1 \text{ and } f(x) = 0 \text{ otherwise,}$$

where θ is a parameter restricted to satisfy $-1 \le \theta \le 1$.

(a) Sketch well-labeled graphs of this function when $\theta = 1$, when $\theta = 0$, and when $\theta = -1/2$.
(b) Verify that for any value of θ satisfying $-1 \le \theta \le 1$, the total area under the density curve does equal one.
(c) Explain why this function does not produce a legitimate probability model for values of θ not satisfying $-1 \le \theta \le 1$. [*Hint*: Drawing some sketches of the function for values of θ outside of that interval might be helpful.]
(d) Evaluate $f(0)$. Does this represent the probability of X equaling zero? Explain.
(e) Determine the expected value μ of this model in terms of θ. [*Hint*: Refer to Exercise 46 for the definition of expected value of a continuous probability model.]

49. Uniform Models

A uniform probability model is one whose probability density function is constant (flat) between two endpoints. Let's call the endpoints a and b, where $a < b$. So the pdf has the form $f(x) = k$ when $a \le x \le b$, 0 otherwise, where k is the appropriate constant. For example, the times at which calls are made to a computer help line in a particular hour period could follow a uniform distribution (0, 60) if they are equally likely to occur at any time in that hour period.

(a) Sketch and label a general uniform(a, b) distribution pdf and determine the constant k, as a function of a and b, so that the total area under the density equals one.
(b) Use integration to determine the expected value μ of the uniform distribution. [*Hint*: Refer to Exercise 46 for the definition of expected value of a continuous probability model.]
(c) Interpret this value geometrically (in other words, where in the interval from a to b does the mean value fall). Explain why this makes sense.
(d) It can be shown that the standard deviation of this uniform distribution is the square root of $(b-a)^2/12$. Determine the standard deviation of a uniform distribution on the interval (0, 2), on the interval (0, 10), and on the interval (8, 10).
(e) Explain why the relative values of these three standard deviations make sense.

50. House Prices

Cal Poly students Peter Cerussi and Patrick Ziegler were interested in studying factors that are related to the price of a house. They gathered data from realestate.com on the listed prices of houses for sale in San Luis Obispo, California on November 20, 2003. The prices of eight houses are shown below, and are in the houseprices.xls Excel file.

Price (in $K): 255, 349, 399, 460, 545, 649, 799, 1195

You will now consider other criteria based on the absolute deviations between the data values and your guess. Even if you keep absolute deviations as your basis for a minimization criterion, you can consider functions other than the sum. For example, if you want to be sure that you are never too far off, you might want to minimize the *maximum* of those absolute deviations:

$$MAXAD(m) = \max\{|255 - m|, |349 - m|, |399 - m|, |469 - m|, |545 - m|, |649 - m|, |799 - m|, |1195 - m|\}.$$

(a) Use the Excel file to investigate the behavior of this *MAXAD* function. Return the data values (house prices) in column A to their original values, and click on cell E2. Notice that this cell contains a formula for evaluating the *MAXAD* function. Use the "fill down" feature to evaluate this function for the rest of the *m* values. Then use Excel to draw a graph of the *MAXAD* function. Describe its behavior, and comment on whether it has a unique minimum value. Identify where the minimum occurs and what that minimum value is.

(b) Change the maximum house price from 1195 to 895 thousand dollars. Comment on the impact of this change on the *MAXAD* function and especially on where the function is minimized.

(c) Change the fourth house's price from 469 to 529 thousand dollars, and reevaluate the *MAXAD* function. Now what has changed, and what has not?

(d) Now change the cheapest house's price from 255 to 305 thousand dollars, and reevaluate the *MAXAD* function. Now what has changed and what has not?

(e) Based on this analysis, make a conjecture for determining the value that will minimize the maximum of absolute deviations from the mean of the data values.

51. House Prices (cont.)

Reconsider the previous Exercise and the houseprices.xls Excel file.

Consider a measure of spread based on absolute deviations: minimizing the *median* of them. Let the function *MEDAD* be defined as:

$$MEDAD(m) = \text{median}\{|255 - m|, |349 - m|, |399 - m|, |469 - m|, |545 - m|, |649 - m|, |799 - m|, |1195 - m|\}.$$

Use Excel to investigate the behavior of this *MEDAD* function. In particular, describe its shape, identify where the function is minimized for the house prices data, and comment on the effects of changing the maximum, middle, and minimum values on the function.

52. House Prices (cont.)

Reconsider the previous Exercise and the houseprices.xls Excel file. You have already investigated finding a prediction that minimizes the sum of absolute deviations and the sum of squared deviations. With the benefit of technology, we need not limit ourselves to exponents of 1 and 2, however. Use technology to examine the function *SkD*(*m*), defined as:

$$SkD(m) = \sum_{i=1}^{n} |x_i - m|^k$$

(a) First analyze this function where $k = 1.5$. Look at a sketch of the function and describe its shape. What value of *m* minimizes this function? Is this minimum value between those for when $k = 1$ and when $k = 2$ (the median and mean, respectively, as you found above)?

(b) Choose another value of *k*, repeat this analysis, and report on your results.

53. Memorizing Letters (cont.)

Reconsider the data from the memory experiment (MemoryLetters.txt)

(a) Use technology to simulate a randomization test to investigate whether the difference in group means is significant. Use at least 1000 repetitions, and report the approximate p-value. Include your technology output and graphical display of the empirical randomization distribution.

(b) Summarize your conclusion and explain how it follows from your simulation analysis. Also address the issue of whether a cause-and-effect conclusion is warranted, paying attention to the design of the study.

(c) Repeat this analysis on the group *medians*, and comment on whether your conclusion differs substantially.

54. Bumpus Data (cont.)

Reconsider the Bumpus data on humerus lengths (Bumpus.txt).

(a) Use technology to simulate a randomization test to investigate whether the difference in group means is significant. Use at least 1000 repetitions, and report the approximate p-value. Include your technology output and graphical display of the empirical randomization distribution.

(b) Summarize your conclusion and explain how it follows from your simulation analysis. Also address the issue of whether a cause-and-effect conclusion is warranted, paying attention to the design of the study.

(c) Repeat (a) and (b) with an analysis of the sparrows' weights.

55. Sleeping Student (cont.)

Reconsider the students' sleeping times from exercise 15 (SleepStudents.txt).

(a) Choose one of these three students, and conduct a simulation analysis to approximate a randomization test for comparing her school night *sleeping times* to her non-school night *sleeping times*. Submit a well-labeled histogram of your simulation results.

(b) Report the approximate p-value based on your simulation results. Does your analysis suggest that the difference in their mean *sleeping times* between school nights and non-school nights is unlikely to have occurred by chance?

(c) Is this a study for which the randomization in your simulation mirrors that in the design, or is the randomization hypothetical in this study? Explain.

56. Musical Dining

A study by North and Shilcock involved three weeks monitoring the effects of classical, pop music, and background silence on customers' spending in British restaurants. Each type of music was played for 6 nights (the order was randomly determined to guard against confounding). When classical music was played in the background, 120 diners spent an average of £24.13 per person on food and drinks. When pop music was played, the 142 diners spent an average of £21.92.

(a) What additional information from the two samples would you need in order to decide if the difference in spending between the classical and pop music was statistically significant?

(b) Sketch comparative boxplots for hypothetical spending distributions between these two groups, creating a situation where you think the difference would be statistically significant. Explain the reasoning behind your sketch.

(c) Sketch comparative boxplots for hypothetical spending distributions between these two groups, creating a situation where you think the difference would *not* be statistically significant. Explain the reasoning behind your sketch.

57. Mirrors and Exercises

In a study reported in the journal *Health Psychology* (Ginis, Jung, and Gauvin, 2003), researchers investigated whether the presence or absence of mirrors during an exercise session would affect women's attitudes toward the session. The subjects were 58 sedentary women, who rode a stationary exercise bike for a 20-minute session. A week later the women returned for another 20-minute session, for which they were randomly assigned to exercise in front of either a mirrored or curtained wall. The first table in the research article describes characteristics of the sample, including the variables of age, body mass index, smoking status, and student status. Some of the statistics reported include:

	Mirror (n_m = 28)			Curtain (n_c = 30)		
Variable	**Proportion**	**Mean**	**Std. Dev.**	**Proportion**	**Mean**	**Std. Dev.**
Smoking	0.071			0.067		
Student	0.786			0.734		
Age		20.86	1.65		20.60	1.57
Body mass index		23.35	3.76		24.23	6.19

(a) Classify each of these four variables as categorical or quantitative.
(b) For each of these four variables, conduct a test of whether the two groups differ significantly on that variable. Report all of the test statistics and p-values. [*Hint*: Check technical conditions as much as possible for all four tests. For testing the proportions, if the technical conditions of the *z*-test are not satisfied, apply Fisher's Exact Test.]
(c) Why do you think the researchers collected and examined these data, performed these tests, and presented the results in the article?
(d) Do you think the researchers were pleased that none of these differences turned out to be statistically significant? Explain why.

58. Fish Oil

Researchers randomly assigned 14 male volunteers with high blood pressure to one of two diets for four weeks: a fish oil diet and regular oil diet. The subjects' diastolic blood pressure was measured at the beginning and end of the study, and the reduction was recorded for each subject (taken from Ramsey and Schafer (2002) based on a study by Knapp and Fitzgerald (1989)). Prior to conducting the study, researchers conjectured that those on the fish oil diet would tend to experience greater reductions in blood pressure than those on the regular oil diet. The resulting reductions in diastolic blood pressure, in millimeters of mercury were

Fish oil diet	8	12	10	14	2	0	0
Regular oil diet	–6	0	1	2	–3	–4	2

(a) Is this an observational study or an experiment? Explain.
(b) Identify the explanatory variable and the response variable. Classify each as categorical or quantitative.
(c) State the hypotheses, in symbols and in words, for testing the researchers' conjecture about this study.
(d) Carry out a randomization test to determine whether the difference in group means is statistical significant at the 0.05 level.
(e) Is it appropriate and valid to carry out a pooled two-sample *t*-test here? Explain.
(f) Conduct a pooled two-sample *t*-test (whether you think it's valid to do so or not). Report the test statistic and p-value. How does the p-value from this pooled *t*-test compare to the p-value from the randomization test in (d)? Would you say that the pooled *t*-test provides a reasonably close approximation to the randomization test in this case? Explain.
(g) Use the pooled *t*-procedure to construct a 95% confidence interval for the treatment effect of the fish oil diet compared to the regular oil diet.

59. Fish Oil (cont.)
Reconsider the previous question about the fish oil study.
(a) Conduct a (non-pooled) two-sample *t*-test and confidence interval. Comment on how the results differ from those of the pooled test. Does the pooling appear to make much difference in this case?
(b) Explain why it would definitely not be appropriate to conduct a paired *t*-test on these data.

60. Fish Oil (cont.)
Reconsider the fish oil study again. Comment on how the p-value from the pooled *t*-test would change in the following situations. Provide an intuitive explanation for your reasoning in each case. Also provide an algebraic explanation based on the test statistic calculation in each case.
(a) What if the group means had been closer together (and everything else had been the same)?
(b) What if the group means had been further apart (and everything else had been the same)?
(c) What if the sample sizes had been larger (and everything else had been the same)?
(d) What if there had been more variability in each sample (and everything else had been the same)?

61. Fish Oil (cont.)
Reconsider the previous question. Comment on how the *width of a confidence interval* for the treatment effect would change in each of the four situations (group means closer together, group means further apart, large sample sizes, more variability in each sample). Again provide both an intuitive and an algebraic explanation for your reasoning in each case.

62. Musical Dining (cont.)
Consider the study by North and Shilcock (Investigation 4.5) where they spent three weeks monitoring the effects of classical, pop music, and background silence on spending. Each type of music was played for 6 nights (the order was randomly determined to guard against confounding). When classical music was played in the background, 120 diners spent an average of £24.13 per head on food and drinks. When pop music was played, the 142 diners spent an average of £21.92.
(a) What additional information from the two samples would you need in order to decide if whether difference in spending between the classical and pop music was statistically significant?
(b) Sketch comparative boxplots for these two groups where you think the difference would be statistically significant.
(c) Sketch comparative boxplots for these two groups where you think the difference would not be statistically significant.

63. Backpack Weights (cont.)
Reconsider the backpack data from the Chapter 1 Exercises (`backpack.txt`). Analyze the data to examine whether the data suggest that male and female students differ significantly with regard to any of three variables: body weight, backpack weight, and ratio of backpack weight to body weight. Include both descriptive (graphical and numerical) and inferential (significance test and confidence interval) components to your analyses. For each variable, write a paragraph or two summarizing your findings.

64. Used Hondas
The `HondasUsed.txt` file contains data on a sample of 16 used Honda Civics for sale on the web on December 7, 2004 and an independent sample of 24 used Honda Accords for sale on the web on December 13, 2004.

(a) Produce graphical displays and numerical summaries to compare the distributions of prices between the two models of cars. Comment on what this descriptive analysis reveals. (Like always, comment on shape, center, spread, and unusual observations.)
(b) Consider these for now to be random samples from the populations of all used Civics and all used Accords for sale on the web in December 2004. Conduct a two-sample *t*-test of the conjecture that Accords tend to cost more on average than Civics. Report your findings. (Include all components of the test in your report.)
(c) Estimate the difference in population means between the two car models with 90%, 95%, and 99% confidence intervals.
(d) Based on these intervals, how confident would you feel about concluding that used Accords cost more than $2000 more on average than used Civics? What about concluding that used Accords cost more than $3000 more on average than used Civics? Explain.

65. Ideal Age

Social scientists have noted that American culture celebrates youth, and they have studied what American consider to be the ideal age. The Harris Poll asked a nationwide sample of 2306 adults on September 16-23, 2003 the following question: "If you could stop time and live forever in good health at a particular age, what age would you like to live at?" The mean response from men was 39 years, and the mean response from women was 43 years.

(a) Consider testing whether this difference in mean responses is statistically significant. What further information would you need to conduct a two-sample *t*-test?
(b) Suppose that the sample sizes were roughly the same for men and for women, so roughly 1153 in each group. With those sample sizes, does the distribution of "ideal ages" need to be normal in order for the *t*-procedures to be valid?
(c) Suppose that in each group, the standard deviation of the "ideal age" responses is 10 years. Sketch the sampling distribution of the test statistic and determine the observed test statistic and *p*-value of the two-sample *t*-test. Is the difference in mean responses significant at the 0.01 level?
(d) Repeat (c) if the standard deviation of the "ideal age" responses is 20 years for each group.
(e) How large would the standard deviation need to be in order for the sample results not to be statistically significant at the .01 level?
(f) Does this (your answer to part e) seem like a reasonable value for the standard deviation in this case? Explain.

66. Health Club Ages

A student collected data on ages of people who joined a local health club in August and September of 2004, also recording the gender of each person (Schmitt, 2004). The student took a systematic sample of every 5th male from a computerized list of males who joined each month and then again for females. The student wanted to test whether the ages of males and females differ significantly and whether the ages of new members in the two months differ significantly. The data are in the file GymMembership.txt.

(a) Start with the question of whether men's and women's ages differ significantly on average. Analyze the data to address this issue. Include both descriptive (graphical and numerical) and inferential (significance test and confidence interval) aspects to your analysis. Include all components (including a check of technical conditions), and summarize your findings.
(b) Repeat (a) for the question of whether mean ages of new members differed significantly between August and September.
(c) To what populations would you feel comfortable generalizing your findings? Explain.

67. Melting Chips

A study was carried out to see whether there is a difference in the melting times of semisweet chocolate chips and peanut butter chips. Twenty students in a statistics class were told to put a chip on their tongue, touch it to the roof of their mouth, and then time how long it was before the chip was completely melted, without any "encouragement" on their part. Each student repeated this with both types of chips, randomly determining which chip they would use first. The data are in the file `ChipMelting.txt`.

(a) Is this an observational study or an experiment? Identify the observational/experimental units and the variables of interest.
(b) Produce graphical displays and numerical summaries to analyze the differences in melting times between the two kinds of chips. Write a paragraph summarizing your findings.
(c) Conduct a two-sample *t*-test to determine whether there is a significant difference in the average melting time between these types of chips. Report the hypotheses, test statistic, and p-value.
(d) Explain why the analysis in (c) is not valid.
(e) Conduct a matched-pairs *t*-test of whether the data suggest that either type of chips tends to take longer to melt than the other. Report all components of the test, including graphical and numerical summaries of the *differences*, the check of technical conditions, and summarize your conclusions. Be sure to comment on whether a cause and effect conclusion can be drawn and the population that you are willing to generalize these results to.
(f) Construct and interpret a 90% confidence interval for the treatment effect on melting time of chocolate as opposed to peanut butter chips.
(g) Suppose that you had calculated the differences in melting times by subtracting in the opposite order. Describe specifically what effect this would have on the test statistic, the p-value, and the confidence interval.

68. Presidential Doctors?

Researchers examined the long-term survival of doctors graduating from one medical school over one century (Redelmeier and Kwong, 2004), comparing those who were presidents of their class to those who appeared alphabetically before or alphabetically after the president in the graduating class photograph. Statistics on long-term mortality were obtained from licensing authorities, medical obituaries, professional associations, alumni records, and national physician directories (follow-up 94%). They reported on 507 presidents and 1014 classmates.

(a) The researchers examined several base-line variables, including gender and whether or not the individual wore glasses. They found 93% of the presidents were male, compared to 85% of their classmates. They also found 9% of presidents were glasses, compare to 12% of their classmates. Are either of these differences statistically significant?
(b) As a measure of accomplishment after graduation, the researchers examined the number of announcements posted by each individual in the alumni notices. They found 21.9% of presidents reported professional accomplishments compared to 13.3% of their classmates. Is this difference statistically significant? (Include all steps of the test of significance and indicate which procedure you are using.)
(c) The overall-life expectancy for the presidents was 49.0 years compared to 51.4 years for their classmates. The two-sided p-value was reported to be 0.036. Assuming the standard deviations were similar in the two samples, use trial-and-error in some technology, or algebra to approximate the value of this standard deviation. What conclusion would you draw from this p-value?
(d) Write a paragraph summarizing your conclusions from these analyses.

69. Exam Performance

Suppose you want to compare student's performances on the first two exams in a course.

(a) Would it make more sense to design this study to use a paired design or an independent sample design? Explain.
(b) For the following summary data, calculate the paired *t*-statistic and *p*-value, and also the independent-samples *t*-statistics and p-value. Does pairing appear to have been useful in this situation? Explain.

Exam 1	$n_1 = 12$	$\bar{x}_1 = 86.4$	$s_1 = 9.5$
Exam 2	$n_2 = 12$	$\bar{x}_2 = 83.3$	$s_2 = 12.3$
Differences	$n_d = 12$	$\bar{x}_d = 3.2$	$s_d = 4.5$

(c) Repeat (b) for the summary data provided next. [*Hint*: If you pay close attention, you can avoid duplicating work.]

Exam 3	$n_3 = 12$	$\bar{x}_1 = 86.4$	$s_3 = 9.5$
Exam 4	$n_4 = 12$	$\bar{x}_4 = 83.3$	$s_4 = 12.3$
Differences	$n_d = 12$	$\bar{x}_d = 3.2$	$s_d = 18.0$

(d) Explain why pairing is so effective in one case and not in the other. You may want to speculate about what else might be true about how students' exam performance is related across the exams.

70. Laptop Fertility

A study published in the on-line journal *Human Reproduction* on December 9, 2004 suggested that using laptop computers could damage the fertility of males by increasing the scrotal temperature, which can affect the quality and quantity of men's sperm. Researchers studied a sample of 29 healthy males between the ages of 21 and 35 by measuring their scrotal temperature before and after using a computer on their laps. The article's abstract reports that the mean temperature was 2.1 degrees Centigrade higher with the computer resting on their laps even when it was not turned on. The mean temperature was 2.7 degrees Centigrade higher with the computer turned on. The abstract of the article did not report standard deviations but said that the p-values were less than 0.0001.

(a) Explain what makes this a matched-pairs design.
(b) State the hypotheses for the appropriate significance test, both in symbols and in words, for comparing temperatures with no laptop and with the laptop turned on.
(c) If the standard deviation of the temperature increases had been 2 degrees Centigrade, calculate the test statistic and p-value.
(d) Determine how large could the standard deviation have been and still produced a p-value of less than 0.0001.
(e) If you had access to the temperature data for each individual subject, what would you examine to assess whether the technical conditions for the paired *t*-test are satisfied? Explain.

71. Seeding Clouds (cont.)

Reconsider the cloud seeding data (`CloudSeeding.txt`) from Investigation 3.9.

(a) Would it be appropriate to carry out a two-sample *t*-test to assess whether the treatment effect is statistically significant? Explain.

When data are strongly skewed, another alternative is to transform data to a scale where the distribution of the variable is more symmetric.

(b) Take the natural log (ln) of each column, and produce graphical and numerical summaries of the distribution of the ln(*rainfall amounts*). Are these distributions now fairly symmetric?
(c) How has the transformation affected the comparison of the standard deviations of each group? Explain why this is also advantageous.
(d) Carry out a two-sample *t*-test on the transformed variable to compare the two groups. Write a paragraph summarizing your conclusion.

72. Heart Transplant Mortality

Reconsider the heart transplant data from Investigation 1.11 (transplants.txt).

(a) Transform the *survival times* by that the natural log. Produce numerical and graphical summaries for comparing ln-survival time between transplant and non-transplant patients. Would it be reasonable to apply the *t*-procedures to these data?

(b) Calculate and interpret a 95% *t*-confidence interval for the difference in mean-ln(*survival times*) between these two groups.

(c) The confidence interval in (b) is for mean(ln(treatment)) – mean(ln(control)).
- Taking into account that the transformed variables are symmetric, how can we express this difference in terms of the medians of the transformed variables?
- Recall from Chapter 3 that the median of the ln-transformed data is equal to the ln of the median of the un-transformed data. Apply this relationship to this difference.
- Now apply a rule of logarithms to this expression to show this confidence interval relates to the ratio of the medians of the un-transformed data.

(d) Exponentiate the endpoints of this interval to obtain a confidence interval for the *ratio* of the group medians.

73. Body Mass Index (cont.)

Reconsider Exercise 37, in which you analyzed body mass index measurements for samples of men and women (BodyMassIndex.txt).

(a) Estimate the difference in mean BMI values between men and women with a 95% confidence interval. Interpret this interval, and comment on whether the technical conditions appear to be satisfied.

(b) Select a transformation that makes the distributions of BMI values more symmetric for both genders. Determine and interpret a 95% confidence interval for the difference in population means on this transformed variable.

(c) Convert your interval in (b) back to the original scale by performing the inverse transformation on the endpoints of the interval. How does this interval compare to the original one in (a)?

74. Freshman Fifteen

Suppose that you want to design a study to investigate the common belief that college freshmen tend to gain fifteen pounds of weight (the so-called "freshman fifteen") during their first term away at college.

(a) Explain why a matched-pairs design would be preferable to a completely randomized design for this study.

(b) State the null and alternative hypotheses, in symbols and in words, for testing this common belief.

(c) Suppose that the mean weight gain during the first term on campus in a random sample of freshmen at one college is 13.6 pounds. What more do you need to know to conduct a matched-pairs *t*-test of your hypotheses?

(d) Describe a scenario in which the sample result in (c) would lead to rejecting the null hypothesis.

(e) Describe a scenario in which the sample result in (c) would lead to not rejecting the null hypothesis.

75. Improving SATs

Suppose that 5000 students are randomly assigned to either take an SAT coaching course or not, with the following results in their improvements in SAT scores:

	Sample Size	**Sample mean**	**Sample std. dev.**
Coaching group	2500	46.2	14.4
Control group	2500	44.4	15.3

(a) Conduct a test of whether the sample data provide evidence that SAT coaching is helpful (in increasing the mean improvement). State the hypotheses, and report the test statistic and p-value. Draw a conclusion in the context of this study.
(b) Produce a 99% CI for the treatment effect of the SAT coaching on improvements. Interpret this interval.
(c) Do the sample data provide very strong evidence that SAT coaching is helpful? Explain whether the p-value or the confidence interval helps you to decide.
(d) Do the sample data provide strong evidence that SAT coaching is very helpful? Explain whether the p-value or the confidence interval helps you to decide.

Exercises 76-78 apply the sign test.

76. Melting Chips (cont.)
Recall the study of chocolate chip and peanut butter chip melting times from Exercise 67 (`ChipMelting.txt`).
(a) Determine whether the semi-sweet chocolate or peanut butter chip melted more quickly, for each student. Record the number of students for which the chocolate chip melted more quickly and the number for which the peanut butter chip melted more quickly. Also construct a bar graph to display these results. [*Minitab Hint*: You could use `MTB> let c3=(c1<c2)` and then `MTB> tally c3`.]
(b) Report the hypotheses, in words and in symbols, for a sign test of whether the data suggest that either type of chip tends to melt more quickly than the other.
(c) Conduct this sign test, report the p-value, and summarize your conclusion.

77. Sleeping Students (cont.)
Reconsider the data from Exercise 15, concerning the nightly sleeping times of college students over a nine-week period (`SleepStudents.txt`). Before analyzing the data, Amber suspected that she tended to sleep longer than either Sarah or Katherine.
(a) For each of the 63 nights, determine who got more sleep between Amber and Sarah (or if they got the same amount of sleep). Construct a bar graph to display the results.
(b) Conduct a sign test of whether the data provide strong evidence that Amber tends to get more sleep than Sarah. Report the hypotheses and p-value, and summarize your conclusion. [*Hint*: First eliminate "ties," nights for which they got the same amount of sleep, from your analysis.]
(c) Repeat (a) and (b) for comparing Amber to Katherine.
(d) If you include ties in the analysis, would it change your findings substantially? Address this question by re-running the sign test, first putting the tie on Amber's "side" and then putting it on Katherine's side. Summarize your findings.

78. Golden Rectangles
The ancient Greeks made extensive use of the "golden rectangle" in art and literature. They believed that a width-to-length ratio of 0.618 was aesthetically pleasing. Some have conjectured that American Indians used the same standard. The following data from Hand et. al. (1994) (also in `shoshoni.txt`) are width-to-length ratios for a sample of 20 beaded rectangles used by the Shoshoni Indians to decorate their leather goods:

0.693	0.662	0.690	0.606	0.570	0.749	0.672	0.628	0.609	0.844
0.654	0.615	0.668	0.601	0.576	0.670	0.606	0.611	0.553	0.933

(a) Produce a histogram and comment on the distribution of these ratios.
(b) Calculate the sample median of these ratios. (Note that the data are not listed in order.)

(c) Conduct a two-sided sign test of whether the sample data suggest that the population median is not 0.618. Report the hypotheses, and show how the p-value is calculated. Also summarize your conclusion.

79. Memorizing Letters (cont.)

Reconsider Exercises 14 and 53 which you analyzed data from a memory experiment.

(a) Analyze these data with a one-sided, two-sample *t*-test. Summarize your findings, including all aspects of a significance test.
(b) Compare your findings to those from an empirical randomization test.

80. Left-Handed Advantages?

Noroozian, Lotfi, Ghassemzadeh, Emami, and Mehrabi (2002) compared the acceptance rate of left-handers with that of right-handers in the College Entrance Examintion (CEE) for the national universities in Iran. About 1 million Iranian high school graduates take part each year in the CEE. An entrance exam score is obtained for each participant, which has a mean of 5000 and a standard deviation of 100. A comprehensive list of all participants between 1993–1997 was obtained, and 10,000 were chosen randomly from each year. Hand preference was exclusively defined as writing preference. The distribution of left-handers and the distribution of right-handers did not differ significantly with respect to gender. Of the 47,854 right-handers, the mean score on the CEE was 5020, with standard deviation 718. Of the 3,398 left-handers, the mean score on the CEE was 5060, with standard deviation 720.

(a) Is it appropriate to apply the two-sample *t*-procedures to these sample data, or do you not have enough information to decide? Explain.
(b) Is this a statistically significant difference in the mean CEE score between the population of right-handers and the population of left-handers?
(c) Compute a 95% confidence interval for the difference in mean score between the left-handed population and the right-handed population.
(d) Explain how this difference may be considered statistically significant but not practically significant. What is the cause for this?

81. Schizophrenic Twins

Recall the study of the volumes of the hippocampus brain regions of monozygotic twins who are discordant for schizophrenia from Practice Problem 3.11 (`hippocampus.txt`).

(a) Carry out a two-sample *t*-test using these data. What conclusion would you draw about whether the mean hippocampus volumes differ between those affected and those unaffected by schizophrenia?
(b) Explain why this test is inappropriate in light of the way the data were collected.
(c) Compare these results to the ones in Practice Problem 3.11. Does the pairing appear to have been effective? Explain.

82. Close Friends

One of the questions asked of a random sample of adult Americans on the 2004 General Social Survey was:

> *From time to time, most people discuss important matters with other people. Looking back over the last six months - who are the people with whom you discussed matters important to you? Just tell me their first names or initials.*

The interviewer then recorded how many names each person gave, with the person's gender.

(a) The relevant parameter for this study can be symbolized as μmen – μwomen. Describe what this parameter means in this context.
(b) State the appropriate null and alternative hypotheses (in symbols) for testing whether American men and women differ with regard to average number of close friends.

The survey responses are summarized in the following table (and in the datafile CloseFriends.txt):

Number of close friends	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Number of men responses	196	135	108	100	42	40	33	654
Number of women responses	201	146	155	132	86	56	37	813

(c) Use technology to produce graphs for comparing the distribution of number of close friends between men and women. Comment on what the histograms reveal about the shapes of the distributions.
(d) Use technology to determine the sample mean and sample standard deviation of the number of close friends for each sex. Report these with appropriate symbols. Also show how to calculate the sample means by hand from the table above.
(e) Conduct a two-sample *t*-test of the hypotheses from (b). Report the test statistic and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.
(f) Produce a 95% confidence interval for the difference in population means (for the number of close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.
(g) Are the technical conditions for the two-sample *t*-test satisfied here? Explain.
(h) Now conduct a test of whether these sample data suggest that the *proportion* of Americans who say they have *zero* close friends differs between men and women. Report the hypotheses, test statistic, and p-value. State your test decision at the 0.05 significance level, and summarize your conclusion.
(i) Produce a 95% confidence interval for the difference in population proportions (who have zero close friends) between men and women. Also write a sentence or two interpreting what the interval reveals.

83. Facebook Emotions

Kramer, Guillory, and Hancock (2014) examined data from Facebook on whether posters respond differently depending on the level of emotional content expressed in the News Feed of their friends. The Facebook News Feed filters content to reduce the amount of information presented at once. Facebook uses an algorithm that aims to identify the content that is most relevant and interesting. In this study, Facebook manipulated how much positive and how much negative content was shown in the feed (to people who viewed Facebook in English). In one part of the experiment, the exposure to positive emotional content was reduced, and in the other the exposure to negative emotional content was reduced. Both studies included a control condition in which a similar proportion of posts were omitted at random. The experiments took place for 1 week (January 11–18, 2012). Participants were randomly selected based on their User ID, resulting in a total of approximately155,000 participants per condition (689,003 participants overall) who posted at least one status update during the experimental period. One response variable was the percentage of all words produced by a person that was either positive or negative during the experimental period.

(a) Below are graphs from this study. Write a few sentences describing what these graphs reveal and whether they appear to support the theory that "affective states are contagious"?

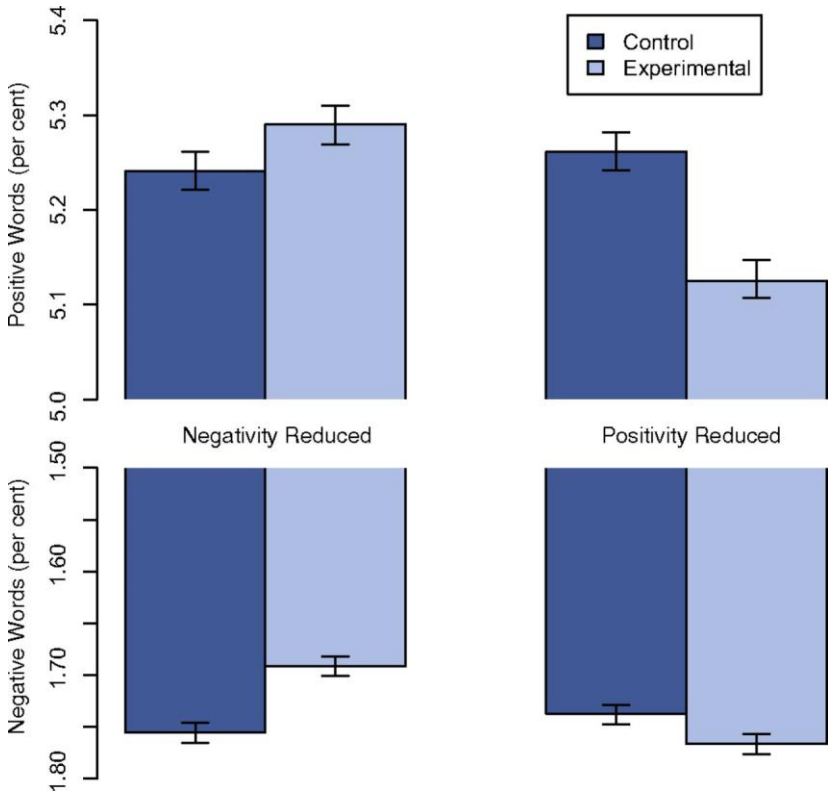


Fig. 1: Mean number of positive (*Upper*) and negative (*Lower*) emotion words (percent) generated people, by condition. Bars represent standard errors.

(b) Consider the mean percentage of positive words generated by people in the negatively reduced condition. Estimate the means and standard errors from the graph and estimate the p-value using a two-sample *t*-test.
(c) Also use your estimates to construct and interpret a 95% *t*-confidence interval.
(d) Would you consider the results examined in (b) and (c) to be statistically significant? Would you consider them to be practically significant? Explain.
(e) Do you have any issues about the ethnical nature of this study? Explain.

Important notes: It is important to note that a user's full content was always available by viewing a friend's content directly by going to that friend's "wall" or "timeline," rather than via the News Feed. Further, the omitted content may have appeared on prior or subsequent views of the News Feed. Finally, the experiment did not affect any direct messages sent from one user to another.
Posts were determined to be positive or negative if they contained at least one positive or negative word, as defined by Linguistic Inquiry and Word Count software (LIWC2007) word counting system, which correlates with self-reported and physiological measures of well-being, and has been used in prior research on emotional expression.

The following are bootstrapping problems:

84. Honda Civics
The following data pertain to a sample of 16 used Honda Civics advertised for sale on the web on December 7, 2004 (a * indicates that the information was not provided, also found in thee file `HondaCivicUsed.txt`):

ID#	age (years)	mileage	price	ID#	age (years)	mileage	price
1	2	29883	15900	9	2	49620	10988
2	1	13415	15900	10	5	24710	9971
3	2	36592	13900	11	4	81000	8995
4	2	37659	13750	12	7	125369	7888
5	2	59723	11995	13	7	120258	6993
6	2	85246	11990	14	12	107000	3800
7	4	52705	11990	15	16	*	1200
8	4	70702	10995	16	3	64939	*

(a) Examine the sample data on the "age" variable. Would a *t*-procedure be appropriate for these data? Explain.
(a) Use the bootstrap procedure to produce a 95% confidence interval for the median age in the population of all used Honda Civics for sale on the web that day.

85. Fish Oil (cont.)
Reconsider the experiment on fish oil and blood pressure, described in previous exercises.
(a) Estimate the treatment effect with a 95% bootstrap interval (for the difference in group means).
(b) How does this interval compare to the *t*-intervals calculated in Exercises 58 and 59?
(c) Estimate the treatment effect with a 95% bootstrap interval (for the difference in group medians).

Exercises involving Normal Probability Plots:

NPP1. Modeling Australian Births
The file `aussiebirths.txt` contains data on births for 44 babies born in one 24-hour period in Brisbane, Australia (http://www.amstat.org/publications/jse/datasets/babyboom.txt). This was a record high number of births in one day. We want to explore the distribution of *time between births*. Note that the fourth column contains the times of the births (in minutes after midnight).
(a) Use technology to calculate the time between births. Then produce a histogram of the *time between births* variable. Describe the characteristics of this distribution.
(b) Use technology to overlay a normal probability curve on this histogram. Does the normal model do a reasonable job of describing these data?
(c) Now overlay an Exponential curve. Does this probability curve appear to be a better model for these data? Explain. [*Hint*: You may want to change the binning so the first bin starts at zero.]

NPP2. Modeling Australian Births (cont.)
Reconsider the previous exercise.
(a) Produce a normal probability plot of the times between births. Describe how the distribution deviates from normality.
(b) Produce and describe an exponential probability plot of the times between births.
(c) Take a ln transformation of the times. Produce a normal probability plot of these transformed data. Does this plot suggest that a normal model might be appropriate for describing the distribution of the ln of the times? Explain.
(d) Take a square root transformation of the times. Produce a normal probability plot of these transformed data. Does this plot suggest that a normal model might be appropriate for describing the distribution of the square root of the times?

NPP3. Normal Distributions?
Consider the (hypothetical) data in the first three columns of the data file GotNormal.txt.
(a) Produce a histogram for each variable, and describe the shape of each distribution.
(b) For each variable, comment on whether a normal model would seem to be appropriate, based on the histogram.
(c) Construct boxplots of the three variables on the same scale. [*Hint*: Using the multiple Y's option.] Describe what these boxplots reveal about similarities and differences among these three distributions and about the appropriateness of the normal model.
(d) For each variable, produce a normal probability plot. Comment on what these plots reveal about the appropriateness of a normal model for each variable. In particular, use these plots to describe *how* the non-normal distributions deviation from the expected behavior of a normal distribution.

NPP4. Modeling Australian births (cont.)
Reconsider the previous exercise. Suppose we think that the times between births in a Australian hospital are well modeled by an exponential distribution with parameter $\beta = 33$ minutes and you want to determine the probability of more than 1 hour (60 minutes) transpiring between births.
(a) Write the function for the density curve with this value of β.
(b) Integrate this function to determine $P(X \geq 60)$.
(c) Use technology to confirm your calculation (scale = 33, threshold = 0).
(d) Use technology to determine how many of the 43 observed times between births were longer than 60 minutes. How does this relative frequency compare to the probability predicted by the exponential model?

NPP5. Body Mass Index (cont.)
The data in BodyMassIndex.txt are ages (in years), weights (in kg), and heights (in cm) for a sample of adults (Heinz et al., 2003). Body mass index (BMI) is defined to be a person's weight (in kg) divided by the square of their height (in meters).
(a) Examine separate normal probability plots for the BMI values of men and women. Does the normal model appear to be appropriate for either sex? For which sex does it come closer to providing a reasonable model? [*Hint*: You may first need to recalculate the BMI values from the weights and heights.]
(b) Try several transformations (log, square root, reciprocal) of the BMI values for the two genders separately. With each transformation, examine separate normal probability plots for men and women. For each gender, identify which transformation produces an approximately normal distribution.

NPP6. 2004 U.S. Open (cont.)
Recall the US Open data (USOpen04.txt). Suppose we wanted to compare whether men tend to take longer to play tennis than women.
(a) Explain why it would not be reasonable to compare the times of the matches for the men and women directly.
(b) Create a new variable "time per set". Examine visual displays, including normal probability plots and boxplots, for comparing the distributions of time per set between men and women. Does the normality condition appear to be met for a two-sample *t* test to compare the average time per set for men and women? Explain.
(c) Take the log of the time per set variable. Would the normality conditions for the two-sample *t*-procedures appear to be met using this as the response variable? Explain.
(d) Carry out the two-sample *t*-test for both the transformed and the untransformed data, testing whether

the data suggest that men tend to take longer to play a set than women do. How do the p-values compare?

(e) Identify the match that is marked as an outlier by the boxplots. Predict how the p-values will change if this match is removed from the analysis.

(f) Investigate your conjecture in (e) by removing that match and reanalyzing both the transformed and untransformed data.

(g) Of the above analyses, which would you consider most appropriate? How would you interpret the p-value in that analysis?

NPP7. Hypothetical Waiting Times

Suppose the data in HypoWaitTimes.txt represent the amount of time patients waited in an emergency room prior to seeing a doctor (in minutes).

(a) Produce numerical and graphical summaries of this distribution and describe what they reveal.

(a) Two different models that are often used to describe waiting times and other skewed right distributions are the "Weibull" density function and the "Lognormal" density function.

(b) Add a "Distribution Fit" to your histogram using the Weibull distribution and then the Lognormal distribution. Comment on the behavior of these models.

(c) Use probability plots to determine whether these data are better modeled by a Weibull density function or a Lognormal density function. Justify your conclusion.

(d) d) For the distribution you choose in (b), use the parameter estimates reported by technology and estimate the probability that a randomly selected person would have to wait more than 240 minutes at this hospital for this fitted distribution.

(e) Use this same distribution to estimate the 90th percentile of waiting times at this hospital.

NPP8. Stock Prices

The file StockchangesOct31.txt contains the opening prices and net changes on October 31, 2001 for 3561 stocks listed on the New York Stock Exchange (nyse.com).

(a) Examine a histogram and boxplot of the opening prices. What unusual feature of this distribution is immediately apparent?

(b) Identify (by its stock exchange symbol) the stock with the largest opening price.

(c) Remove this outlier from the analysis, and then produce a histogram and boxplot of the remaining prices. Is there still an outlier that dominates these graphs? If so, identify its stock market symbol.

(d) Remove this second outlier from the analysis, and then produce a histogram and boxplot of the opening prices. Describe the distribution of opening prices now that two outliers have been removed.

(e) Examine visual displays and describe the distribution of net changes, leaving those two outlying stocks out of the analysis.

(f) Examine normal probability plots of the opening prices and net changes. Does the normal model seem to be appropriate for either variable? If not, describe how the distribution(s) deviates from normality.

(g) What percentage of the net changes fall within 1 standard deviation of the mean? Does this provide further evidence about the suitability of the normal model? Explain.

(h) Create a new variable: *percentage change*. [*Hint*: Divide the net change by the opening price and multiply by 100.] Examine visual displays, including a normal probability plot. Comment on its distribution, including whether the normal model would be appropriate for describing these percentage changes.

(i) For BRK A, the opening price was $69,800 and the net change was $1200. Calculate the percentage change for this stock. Is this percentage more than 2 standard deviations from the mean percentage change for the data set in (e)? If not, explain how this stock could be such an extreme outlier in terms

of net change, but not percentage change.

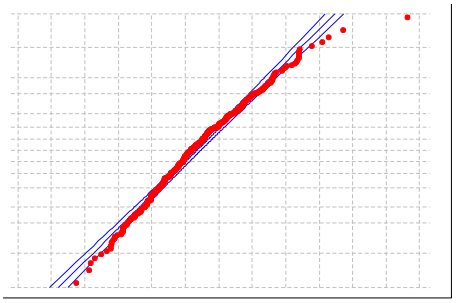
(j) Repeat (i) for BRK B.

Note: Typically, when a stock price rises enough, a company will "split" the stock (each new share is worth half the value of the old shares), believing these lower-priced shares will be more attractive to investors. BRK A is the Berkshire Hathaway stock (class A) and BRK B is the Berkshire Hathaway stock (class B). Berkshire Hathaway is run by Warren Buffet, the "oracle of Omaha," who does not believe in stock splits, so the price of shares of these stocks has increased over time while other stocks increasing in value have generally split.

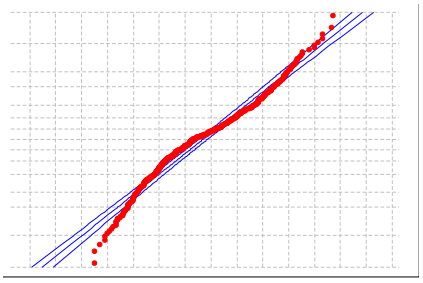
NPP9. Matching Probability Plots to Boxplots

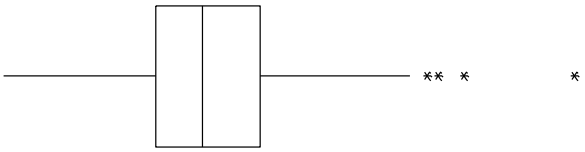
Graphs for three different variables are given below, one boxplot and one normal probability plot for each. Which boxplot corresponds to which normal probability plot? Write a few sentences providing your justification.

a) I

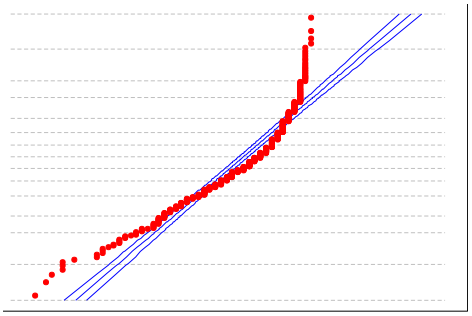


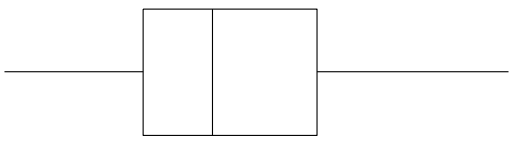
(b) II





(c) III





The following are one-mean *t*-tests/confidence interval exercises:

OST1. Academy Award Mortality
Redelmeier and Singh (2001) wanted to see whether the increase in status from winning an Academy Award is associated with long-term mortality among actors and actresses. They found 235 actors and actresses who had won at least one academy award and 527 who had been nominated but never win. At the time of the analysis, the average life expectancy for the winners was 79.7 years compared to 75.8 years for the nominees.
(a) Assuming a standard deviation of 17 years for each group, is this a statistically significant difference?
(b) In identifying the actors and actresses for this study, nominated actors/actresses were paired with a person of the same gender and similar age from the same film. Explain the advantages of this pairing.
(c) Summarize the conclusions you would draw from this study.

OST2. Smoking Habits
One of the questions in the National Health and Nutrition Examination Surveys (NHANES) study asked subjects about their smoking habits. One of the questions asked whether the person has smoked at least 100 cigarettes in his/her life. The 2328 people who answered "yes" were asked to report the age at which they started smoking. The responses are tallied in the table below and in the file SmokingStart.txt:

Age	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Count	10	6	10	23	24	99	115	155	255	195	239	377	152	192	120
Age	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**
Count	72	40	29	64	20	8	17	10	36	2	1	3	4	15	2
Age	**37**	**38**	**40**	**41**	**43**	**45**	**46**	**47**	**49**	**50**	**54**	**55**	**65**	**72**	
Count	4	2	9	2	4	3	1	2	1	1	1	1	1	1	

For now, consider these 2328 smokers to constitute the entire population of interest.
(a) Examine visual displays (histogram, boxplot) of the distribution of ages, and write a paragraph summarizing its features.
(b) Report the mean and median, standard deviation and IQR of these ages. Are these parameters or statistics? What symbols would you use for the mean and standard deviation?
(c) Suppose that we were to take a simple random sample of 40 people from this population of 2328 smokers. Would you expect the sample mean age to equal the population mean exactly? Explain.
(d) Does the Central Limit Theorem for a sample mean apply in this case? In other words, can the CLT tell us about the sampling distribution of the sample mean age if we were to repeatedly take random samples of size 40 from this population? If not, explain. If so, describe what it says in this case, and draw a well-labeled sketch of the sampling distribution.
(e) According to the CLT, what is the probability that the sample mean age of 40 randomly selected people from this population would exceed 20 years? (Show the details of your calculation and/or relevant output from technology.) Shade the region of interest on your sketch, and write a one-sentence summary of the probability.
(f) According to the CLT, what is the probability that the sample mean age would be less than 17.5 years? (Show the details of your calculation and/or relevant output from technology.)
(g) According to the CLT, what is the probability that the sample mean age would fall between 18 and 19 years? (Show the details of your calculation and/or relevant output from technology.)

OST3. Smoking Habits (cont.)
Reconsider the previous question about ages at which people started to smoke. Continue to regard those 2328 smokers as the entire population of interest, and consider taking a random sample of 40 smokers.
(a) Write and execute a simulation for taking 1000 random samples of size 40 from this population,

recording the sample mean age for each sample. Construct a histogram and calculate descriptive statistics for the 1000 sample mean ages.
(b) Are your findings in (a) close to what the CLT would predict? Explain.
(c) Use your simulation results to approximate the probabilities asked for in (e)–(g) of the previous question. Comment on how closely the simulation results match those from the CLT.

OST4. Smoking Habits (cont.)
Reconsider the data on ages at which people start smoking, but now consider the 2328 smokers to be a random sample from the population of all smokers in the U.S.
(a) Use the sample data to conduct a significance test of whether the mean age at which smokers begin to smoke differs from 18 years. Report the hypotheses in symbols and in words, comment on the technical conditions, and calculate the test statistic and p-value. Include a well-labeled sketch of the sampling distribution for the test statistic and indicate the area represented by the p-value. Also indicate whether the sample mean differs significantly from 18 at the 0.10 level, the 0.05 level, and the 0.01 level. Summarize your conclusions.
(b) Construct and interpret a 95% confidence interval for the population mean age at which smokers begin to smoke.
(c) Do you expect that about 95% of the ages in this sample fall within this interval? Would you expect that about 95% of the ages in the population of American smokers fall within this interval? Explain.
(d) Would you consider it to be valid to calculate a prediction interval with these data? If not, explain. If so, do so and interpret your result.

OST5. Margin-of-Error Properties

Consider the margin-of-error for a *t*-interval estimating a population mean μ: $t^* \frac{s}{\sqrt{n}}$.

(a) Explain what each of these symbols (t^*, s, n) represents.
(b) Is the margin-of-error an increasing or decreasing function of t^*, or is it neither? Is it an increasing or decreasing function of the confidence level? Explain both mathematically and intuitively.
(c) Is the margin-of-error an increasing or decreasing function of s, or is it neither? Explain both mathematically and intuitively.
(d) Is the margin-of-error an increasing or decreasing function of n, or is it neither? Explain both mathematically and intuitively.
(e) Does doubling the sample size cut the margin-of-error in half, if everything else remains the same? Explain.

OST6. Margin-of-Error Properties (cont.)

Reconsider the margin-of-error for a *t*-interval estimating a population mean μ: $t^* \frac{s}{\sqrt{n}}$. Suppose that you want to determine the sample size n needed for the margin-of-error not to exceed some pre-specified bound, M, at a certain confidence level.
(a) Solve for an inequality expressing the necessary sample size n as a function of t^*, s, and the error bound, M.
(b) Is this an increasing or decreasing function of t^*? Of the confidence level? Of s? Of M? Explain why your answers make intuitive sense.

OST7. Backpack Weights
Recall the data on backpack weights from Exercise 63 in Chapter 1. Now consider the backpack weights themselves as a ratio to the individuals' body weights (backpack.txt).
(a) Construct graphical and numerical summaries to describe the distribution of the weight ratios. Comment on what this preliminary analysis reveals.
(b) Conduct a significance test of whether the sample data suggest that the mean weight ratio among all Cal Poly students is actually less than 0.10. Report hypotheses, comment on technical conditions, and calculate the test statistic and p-value. Include a well-labeled sketch of the sampling distribution for the test statistic and indicate the area represented by the p-value. Also summarize your conclusion and explain how it follows from your test.
(c) Construct and interpret a 90% confidence interval for the population mean of the weight ratios.
(d) Do you have any concerns about sampling bias or non-sampling errors with this study? Explain.

OST8. Honda Civics
Recall the data on used Honda Civics from Exercise 82 (HondaCivicUsed.txt).
(a) Identify the observational units with these data.
(b) Identify the five variables represented here (the *model* is not a variable here). Identify each as categorical or quantitative.
(c) Examine graphical displays and numerical summaries for the age, mileage, and price variables. Comment on the distribution of each variable in this sample.
(d) Treat these as a random sample from the population of all used Honda Civics for sale on the web that day. Would you feel comfortable applying a *t*-interval to estimate the population mean for any of these variables? For all of them? Explain.
(e) Construct and interpret a 95% confidence interval for the population mean price of used Honda Civics for sale on the web.
(f) Would you consider it appropriate to use these data to construct a prediction interval for the price of an individual Honda Civic for sale on the web that day? If not, explain. If so, construct and interpret this prediction interval.
(g) How large a sample would be needed to estimate the population mean price to within ±500 dollars with 90% confidence? (Use the standard deviation of prices in this sample as your best estimate of the population standard deviation.)
(h) Is there any sample size for which the half-width of a 90% prediction interval for price would be 500 dollars or less? Explain.

OST9. Breaking Ice
Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file NenanaIceBreak.txt contains all of the data since 1917.
(a) Examine and comment on graphical displays of the "date" variable, recorded in days with April 1 being coded as 1. [*Hint*: Remember to comment on shape, center, and spread, and relate your comments to the context.]
(b) Treat these data as a random sample from the process by which nature produces the ice-breaking dates each year. Produce a 95% confidence interval for the population mean date. Then translate the

endpoints from the coded scale to the actual calendar, and interpret the interval.
(c) Produce a 95% prediction interval for the ice break-up date in an individual year. Again translate the endpoints from the coded scale to the actual calendar, and interpret the interval.
(d) Repeat (a)–(c) for the *time of day* variable with midnight = 0.

OST10. *z* vs. *t*-intervals
Some textbooks recommend that when the sample size is 30 or more, it's ok to use a *z*-interval instead of a *t*-interval, even when you have to estimate the population standard deviation σ with the sample standard deviation *s*, because the intervals do not differ too much. Investigate this recommendation in the $n = 30$ case as follows.
(a) Calculate the widths of a 95% *z*-interval and a 95% *t*-interval (in terms of *s* and *n*). Then calculate the difference in widths and divide by the width of the *t*-interval (the correct one) to determine the percentage error in the width of the *z*-interval.
(b) Use simulation with the Simulating Confidence Intervals applet to compare the coverage rates of the two procedures, assuming that the population follows a normal distribution. (Use at least 1000, preferably 10,000 or more, samples to approximate the coverage rate. Choose at least two different values of the sample size to compare.)
(c) Repeat (b), but with a uniformly distributed population.
(d) Repeat (b), now with an exponentially distributed population.
(e) Summarize your findings.

OST11. Stock Prices
Reconsider the exercise about stock prices (`StockChangesOct31.txt`). Consider the 3559 stocks' opening prices (after removing the two extreme outliers as you did in the previous exercise) to be the entire population of interest.
(a) Is the population distribution symmetric or skewed?
(b) Determine the mean and standard deviation of this population. Record them with the appropriate symbols.
(c) Suppose that you take many random samples of size $n = 5$ stocks from this population and calculate the sample mean for each sample. Would you expect the sampling distribution to be as skewed as the population, less skewed than the population, or nearly symmetric? Explain.
(d) Write a simulation to take 1000 random samples of size $n = 5$ stocks from this population and to calculate the sample mean for each sample. Produce a histogram, boxplot, and normal probability plot of the sample means. Describe this distribution.
(e) Calculate the mean and standard deviation of these 1000 sample means. Are they close to what you would have expected? Explain.
(f) Repeat (b)–(e) with samples of size $n = 40$ stocks. Also comment on how this empirical sampling distribution compares to that when $n = 5$.
(g) Use the Central Limit Theorem to calculate the theoretical probability that a sample mean opening price would exceed 25, with a random sample of size $n = 40$ from this population.
(h) What proportion of your 1000 simulated sample means exceed 25? Is this close to the probability in (g)?

OST12. Stock Prices (cont.)
Reconsider the previous exercise, but turn your attention to the "net change" variable rather than opening price. Repeat (a)–(f) for this variable.